CSD UNITY User Guide

The Cambridge Structural Database in Unity Database Format

2015 CSDS Release

Registered Charity No 800579

Conditions of Use

The Cambridge Structural Database System (CSD System) comprising all or some of the following:

ConQuest, Quest, PreQuest, Mercury, (Mercury CSD and Materials module of Mercury), VISTA, Mogul, IsoStar, SuperStar, web accessible CSD tools and services, WebCSD, CSD Java sketcher, CSD data file, CSD UNITY, CSD MDL, CSD SDfile, CSD data updates, sub files derived from the foregoing data files, documentation and command procedures (each individually a Component) is a database and copyright work belonging to the Cambridge Crystallographic Data Centre (CCDC) and its licensors and all rights are protected. Use of the CSD System is permitted solely in accordance with a valid Licence of Access Agreement and all Components included are proprietary. When a Component is supplied independently of the CSD System its use is subject to the conditions of the separate licence. All persons accessing the CSD System or its Components should make themselves aware of the conditions containe in the Licence of Access Agreement or the relevant licence.

In particular:

- The CSD System and its Components are licensed subject to a time limit for use by a specified organisation at a specified location.
- The CSD System and its Components are to be treated as confidential and may NOT be disclosed or re-distributed in any form, in whole or in part, to any third party.
- Software or data derived from or developed using the CSD System may not be distributed without prior written approval of the CCDC. Such prior approval is also needed for joint projects between academic and for-profit organisations involving use of the CSD System.
- The CSD System and its Components may be used for scientific research, including the design of novel compounds. Results may be published in the scientific literature, but each such publication must include an appropriate citation as indicated in the Schedule to the Licence of Access Agreement and on the CCDC website.
- No representations, warranties, or liabilities are expressed or implied in the supply of the CSD System or its Components by CCDC, its servants or agents, except where such exclusion or limitation is prohibited, void or unenforceable under governing law.

Licences may be obtained from:

Cambridge Crystallographic Data Centre
12 Union Road
Cambridge CB2 1EZ, United Kingdom

Web:	http://www.ccdc.cam.ac.uk
Telephone:	+44-1223-336408
Email:	admin@ccdc.cam.ac.uk

(UNITY is a product of Certara and MDL is a registered trademark of Elsevier MDL)

Contents

1 Introduction

This document describes CSD UNITY, a version of the Cambridge Structural Database (CSD) that can be searched with the UNITY Chemical Information System (TRIPOS, Inc.).

The document describes the procedure needed for installation of CSD UNITY and summarises the main differences in content between CSD UNITY and the standard version of the CSD. Known limitations of this release are referred to.

This document is complementary to the UNITY and CSD System Documentation and should be read in conjunction with them.

2 Installation Instructions

2.1 Restrictions

This version of CSD UNITY was generated under the Debian 4.0 Linux operating system and tested using UNITY8.0. It is not possible to search this database with versions of UNITY earlier than this. If you are running an earlier version of UNITY we recommend you contact Tripos for assistance.

2.2 Copying CSD UNITY to Disk

All CSD UNITY files are tar-ed and gzip-ed in the file `CSD_UNITY.tar.gz`. Users should copy this file and then gunzip and untar it. Files will be restored to a directory called CSD_UNITY. This can be accomplished with the command

```
gunzip -c CSD_UNITY.tar.gz | tar xvf -
```

The resulting `CSD_UNITY` directory contains the files:

```
coord2d.dat    screendefs.dat
coord2d.idx       screendefs.idx
coord3d.dat    Security/
coord3d.idx       sln.dat
dictionary.dat    sln.idx
dictionary.idx
rawsln.dat
rawsln.idx
registry_master.dat
registry_master.idx
registry_part.dat
registry_part.idx3 CSD UNITY
```

2.3 Opening, Selecting and Searching CSD UNITY

Full details of how to open and select UNITY databases may be found in the UNITY User Guide. A brief summary is given here:

1. Select Open from the UNITY File menu and select Database
2. Browse to the directory containing CSD UNITY and select CSD UNITY.
3. Entries in CSD UNITY will be displayed in the UNITY Hitlist Manager window.

Alternatively, CSD UNITY may be accessed and searched through the UNITY menu command of the SYBYL program. The user should refer to the UNITY and SYBYL User Guides for further details.

3 Database Description

3.1 General Description

CSD UNITY is a fully-registered, ready-to-use 2D/3D molecular database in binary form compatible with the UNITY8.0 searching software of TRIPOS, Inc. It was first derived from the Cambridge Structural Database (CSD) version 5.10 (released October 1995) and it has been produced with each subsequent CSD release.

The CSD contains experimentally derived atomic coordinate data from crystal-structure determinations of organic and organometallic compounds undertaken by refinement of X-ray and neutron diffraction data.

3.2 Omitted Entries

The following types of entries were omitted from this version of CSD UNITY:

- Entries containing pi-bonds to metals, e.g. ferrocene (see Chemical Connectivity Descriptions).
- Entries containing crystallographic disorder.
- Entries that are classified in the CSD as being in error.
- Entries that do not have 3D coordinates.
- A small number of entries that, for various reasons, could not successfully be converted to the UNITY format.

3.3 Handling of Mixtures

In the CSD, each entry contains coordinate data for the atoms in the asymmetric unit of the crystallographic unit cell. The asymmetric unit may contain more than one distinct molecular entity, or residue, e.g. an anion/cation pair, a host molecule and its guest, or two chemically-identical but crystallographically independent molecules. In CSD UNITY, each residue has been registered as a separate entry. However, very small residues (e.g. water) have been omitted completely.

3.4 Registration Keys

Each entry in the CSD is identified by a reference code (or REFCODE). The REFCODE comprises 6 alphabetic characters which may be followed by two numeric digits (REFCODnn). A description may be found in the CSD System Documentation. To maintain an identifiable and traceable connection with the original CSD entry, all entries in CSD UNITY are tagged with a key-field derived from the CSD REFCODE. The REFCODE itself could not be used unchanged because many CSD entries contain multiple residues and each residue is registered as a separate entry in CSD UNITY (see Handling of Mixtures). The CSD REFCODE was therefore augmented so that residues deriving from the same CSD entry could be given different registration keys in CSD UNITY. The augmentation involved the addition of two numeric digits to the end of the CSD REFCODE to give a code REFCODnnxx. REFCODnn represents the original CSD REFCODE; if nn was not originally present then two zeros were inserted. xx is used to distinguish individual entries in CSD UNITY that are derived from a single entry in CSD. For example:

- The original CSD entry for *methyl bacteriopheophorbide-a benzene solvate* (CSD REFCODE = BAVSUM01) contains two different residues: methyl bacteriopheophorbide-a and benzene. It was broken into two separate CSD UNITY 5 CSD UNITY entries. Each residue is distinguished by appending to the REFCODE an additional pair of numeric digits (xx), 01 or 02. The resulting CSD UNITY registration keys are thus BAVSUM0101 (for the methyl bacteriopheophorbide-a residue) and BAVSUM0102 (for the benzene residue).
- The original CSD entry for *ß-L-arabinose* (CSD REFCODE = ABINOS) contains a single residue. In CSD UNITY, this becomes a single entry with the registration key ABINOS0001.

3.5 3D Atomic Coordinates

UNITY will only store one set of 3D atomic coordinates for each chemically-unique residue. This is important when the same residue occurs more than once in the CSD, which may happen because:

- The same crystal structure has been determined several times (e.g. by separate research workers or under different experimental conditions), or in two or more polymorphic forms.
- A crystal structure contains two or more crystallographically-independent molecules in the unit cell.
- A residue is found in several different crystal structures (e.g. benzene is a common solvate molecule in organic and organometallic crystal structures).

In each of these cases, the different occurrences of the residue will be separately registered into CSD UNITY, *but only one set of 3D coordinates will be stored* (*viz.* that corresponding to the first-registered entry). 3D coordinates that are not stored in CSD UNITY may be retrieved from the standard export version of CSD using CSD System software.

3.6 Transfer of Metal Complexes to SYBYL

A problem may occur if the user downloads organometallic CSD UNITY entries to SYBYL via the **File... To Sybyl** command of the UNITY list manager. The symptom is that the registration key appears in the SYBYL molecular spreadsheet, but an attempt to load the structure into a SYBYL molecule area produces an error message. The problem may be overcome by typing the instruction

```
   PARAMETER OPEN $TA_DEMO/metals.tpd |
```

at the SYBYL command prompt. Alternatively, such entries can be transferred to SYBYL by writing them out from UNITY as a MOL2 file.

3.7 Chemical Connectivity Descriptions

There are some incompatibilities between the chemical-connectivity conventions used in the CSD and UNITY. First, current versions of UNITY do not recognise pi-bonds to metals (e.g. as in ferrocene). The pi-bonds have therefore been removed from the connectivity representation and, where this has resulted in the pi-ligand being disconnected from the rest of the molecule, the entire entry has been omitted from CSD UNITY. Secondly, there is no equivalent in UNITY of the CSD *polymeric* bond, which is used in the CSD to signify the bond between monomer units in a polymeric structure. In CSD UNITY, polymeric bonds are treated as normal single bonds. Thirdly, there is no equivalent in UNITY of the CSD *delocalised double* bond type, which is sometimes used to represent conjugated systems such as carboxylates and acetylacetonates. In the current version of CSD UNITY, these bonds are treated as double bonds. If in doubt, the safest option is to allow *any bond type* for the potentially ambiguous bonds when constructing the search query.

3.8 Search Queries Containing Hydrogen Atoms

If a search is done for a substructure containing explicit hydrogen atoms, entries will only be found in CSD UNITY if the explicitly specified hydrogen atoms were located in the crystal-structure determination, i.e. database entries with missing H atoms will not be found.

3.9 3D Searching of Stereoisomers

Entries were converted to UNITY format using the UNITY *dbimport* option *+perceive_chiral_C*. In consequence, molecules differing only in the chirality (or chiralities) of one (or several) carbon atom(s) are treated as distinct entries, each with a separate set of 3D coordinates. Thus, a 3D search with distance constraints that should be satisfied by, e.g., ß-D-glucose but not alpha-D-glucose will successfully find the former but not the latter.

Due to processing difficulties, the perception of N/P chirality and E/Z double-bond stereochemistry was turned off, i.e. the *dbimport* options -*perceive_chiral_np* and -*perceive_stereo_bond* were used. 3D searches will not be completely reliable where CSD UNITY contains both isomers of a molecule with a chiral N or P atom, or both the E and Z isomer of a molecule with an asymmetric double bond. Specifically, a 3D search which should find one isomer but not the other will either find both or neither.

3.10 Chemical Diagrams

Chemical diagrams in CSD UNITY are derived algorithmically by the UNITY software. They are not always of the same quality as the hand-drawn diagrams in the CSD.

4 Support

If you have a technical problem with this CCDC product you can contact User Support in Cambridge, who will try to help.

Email:	support@ccdc.cam.ac.uk
Tel:	+44 1223 336022
Fax:	+44 1223 336033
Website:	www.ccdc.cam.ac.uk/

Address:
User Support
Cambridge Crystallographic Data Centre
12 Union Road
Cambridge CB2 1EZ
United Kingdom

If you need to contact User Support, please try to provide the following information:

- The name and version number of the product with which you are having problems.
- The make, model and operating system of the workstation you are using.
- A clear description of the problem and the circumstances under which it occurred.

Also be prepared to email error messages and other output. This information is always useful when trying to determine the cause of a problem.

We try to deal with User Support queries within one working day but sometimes problems can take a little longer to solve. When this happens we will keep you informed of our progress and try to provide you with an answer as quickly as possible.